Bulletin of L.N. Gumilyov Eurasian National University. Mathematics, computer science, mechanics series, 2025, Vol. 150, №1, P. 6-16.
http://bulmathmc.enu.kz, E-mail: vest_math@enu.kz

Статья
МРНТИ: 28.23.17

ИДЕНТИФИКАЦИЯ ЯЗЫКА УСТНОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ WAV2VEC2 ДЛЯ КАЗАХСКОГО ЯЗЫКА [1]

Умбет С. [1], Кожирбаев Ж.М. [2]
[1] *Трирский университет, ул. Университетсринг 15, 54296, Трир, Германия*
[2] *National Laboratory Astana, пр. Кабанбай батыра, 53, Астана, Казахстан*
E-mail:zhanibek.kozhirbayev@nu.edu.kz

Аннотация. В данном исследовании представлена разработка и тонкая настройка модели идентификации языка устной речи с использованием варианта XLSR (Cross-Lingual Speech Recognition) Wav2Vec2. Обученная на богатом и разнообразном наборе данных, охватывающем шесть языков, с особым акцентом на языках с низкими ресурсами, таких, как казахский, модель демонстрирует замечательные возможности в многоязычном распознавании речи. Благодаря обширной оценке тонко настроенная модель не только превосходит существующие эталонные показатели, но и превосходит другие современные модели, включая варианты Whisper. Достигнув впечатляющего результата F1 92,9% и точности 93%, модель демонстрирует свою производительность в реальных многоязычных и малоресурсных сценариях. Эта работа вносит значительный вклад в развитие технологий распознавания речи, предоставляя надежное решение для идентификации языка в различных языковых средах, особенно в недостаточно представленных языковых условиях. Его успех подчеркивает потенциал моделей на основе Wav2Vec2 в улучшении систем обработки речи в многоязычных контекстах с низкими ресурсами. Итоги данного анализа могут способствовать разработке надежных и эффективных систем автоматического распознавания речи, оптимизированных для казахского языка. Такие технологии найдут применение в различных областях, включая преобразование речи в текст, работу голосовых ассистентов и инструменты голосовой коммуникации.

Ключевые слова: идентификации языка, идентификации языка устной речи, казахский язык, Wav2Vec2, XLSR.

DOI: https://doi.org/10.32523/bulmathenu.2025/1.1

2000 Mathematics Subject Classification: 68T10

1. ВВЕДЕНИЕ

Идентификация языка (LID) является важнейшей задачей в области обработки речи, выступая в качестве основы для многочисленных передовых приложений, таких, как автоматическое распознавание речи (ASR), перевод в реальном времени и виртуальные

[1]Работа выполнена при поддержке грантового финансирования проектов Комитета науки Министерства науки и высшего образования Республики Казахстан (грант No. AP23489529)

помощники, которые могут обслуживать многоязычные среды [1]. Способность систем обнаруживать и адаптироваться к разным языкам на лету является не только вопросом удобства, но и необходимостью в сегодняшнем все более глобализированном мире. Однако создание эффективных моделей LID далеко не просто, особенно при работе с языками, которые имеют ограниченные обучающие данные или ресурсы. Недавние достижения в области обработки речи сместились в сторону более сложных, эффективных с точки зрения данных подходов, которые используют кросс-языковое трансферное обучение. Одним из выдающихся методов в этом отношении стала разработка Wav2Vec2, которая произвела революцию в том, как обрабатываются речевые данные, путем обучения представлений непосредственно из необработанных данных [2]. Основываясь на этом, вариант XLSR Wav2Vec2 стал важным решением для задач многоязыковой идентификации языка. Изучая общие речевые представления на разных языках, XLSR оказался особенно эффективным для языков с низкими ресурсами — языков, где доступность обучающих данных ограничена, что затрудняет эффективное обучение моделей с использованием традиционных методов. Основное внимание в этой работе уделяется изучению того, как XLSR можно настроить для решения задач идентификации речи на разных языках. В частности, эта работа нацелена на шесть языков: казахский, русский, английский, итальянский, испанский и корейский. Стоит отметить, что казахский, один из основных языков в этом исследовании, часто недостаточно представлен в наборах данных идентификации языка, что делает его отличным кандидатом для проверки адаптивности модели к средам с низкими ресурсами. Включая сочетание языков с высоким ресурсом (например, английский, русский) и низким ресурсом (например, казахский), это исследование направлено на демонстрацию универсальности и надежности модели в реальной многоязычной среде. В этой работе не только описывается процесс обучения и метрики оценки, но и проводятся сравнения с возможностями LID существующей модели ASR, включая несколько вариантов модели Whisper. Эти сравнения необходимы для объективной оценки сильных сторон и ограничений тонко настроенной модели XLSR. Результаты этого исследования показали многообещающие результаты, особенно с точки зрения точности и оценок F1, что указывает на то, что тонко настроенная модель способна эффективно справляться с задачей идентификации многоязычного языка, особенно для языков с низким ресурсом, таких, как казахский. Основные вклады этого исследования включают:

1. Разработка и настройка модели на основе XLSR: Создание тонко настроенной модели, специально оптимизированной для задачи идентификации языков, включая казахский.
2. Сравнительный анализ с современными моделями: Проведение детального сравнения с существующими ASR-системами, включая несколько вариантов модели Whisper, для объективной оценки производительности.
3. Обоснование возможностей Wav2Vec2 для LID: Подтверждение эффективности использования модели Wav2Vec2 и ее вариантов, таких, как XLSR, для задач многоязычной идентификации речи.

Данная статья организована следующим образом: в разделе 2 представлено подробное обсуждение технических аспектов моделей Wav2Vec2.0 и XLSR, а также обзор литературы, связанной с идентификацией устной речи. Раздел 3 содержит детальное описание использованного набора данных, а также подходов, основанных на архитектуре XLSR, примененных в наших экспериментах. Результаты проведенных экспериментов изложены в разделе 4. Наконец, в разделе 5 подведены итоги исследования, представлены основные выводы, сделанные на основе экспериментов, и выделены перспективные направления для дальнейших исследований.

2. ОСНОВНАЯ ЧАСТЬ

В этом разделе представлен краткий обзор соответствующей литературы, связанной с этой статьей, разделенный на два подраздела: Wav2Vec2.0 и XLSR, а также идентификация языка устной речи.

Wav2Vec 2.0 представляет собой тщательно разработанную модель для расшифровки речи, встроенной в аудиосигналы, с применением самоконтролируемой методологии предварительного обучения. Эта методология позволяет извлекать полезные представления из больших объемов немаркированных аудиоданных. Модель объединяет ключевые принципы нескольких предшествующих подходов, включая Contrastive Predictive Coding (CPC) [3], Model Predictive Control (MPC) [4], wav2vec [5] и vq-wav2vec [6]. Архитектура Wav2Vec 2.0 интегрирует сверточные нейронные сети (CNN) и трансформеры, что позволяет эффективно обрабатывать как локальные особенности, так и глобальные шаблоны в аудиоданных. В основе модели лежит многоуровневый сверточный кодировщик признаков $f : X \to Z$, который преобразует необработанные аудиосигналы X в скрытые представления речи $z_1, \ldots, z_T$. Эти представления затем обрабатываются трансформерной сетью с использованием маскирования $g : Z \to C$, которая преобразует скрытые представления в дискретные выходные данные $q_1, \ldots, q_T$, используемые в качестве целевых значений в задаче самоконтролируемого обучения [7, 8]. Модуль трансформера контекстуализирует квантованные представления с помощью блоков внимания, создавая дискретные контекстуальные представления $c_1, \ldots, c_T$. Кодировщик признаков состоит из семи сверточных блоков с 512 каналами, шириной ядер {10, 3, 3, 3, 3, 2, 2} и шагами {5, 2, 2, 2, 2, 2, 2}. Трансформерная сеть включает 24 блока с размерностью 1024 и внутренней размерностью 4096, а также 16 головок внимания, обеспечивающих высокую степень контекстуализации.

XLSR [9] — многоязычная модель, созданная на основе межъязыковой модели XLM-R, предназначенная для решения задач многоязычной и межъязыковой обработки естественного языка (NLP). Она базируется на архитектуре Wav2Vec 2.0 и обладает уникальной способностью извлекать скрытые квантованные представления, охватывающие множество языков. Это достигается использованием метода квантования произведения, который выбирает квантованные представления из кодовых книг, а процесс выбора реализуется через дифференцируемую технику Gumbel-Softmax. Архитектура XLSR напоминает модель двунаправленного кодировщика представлений трансформера (BERT) [8], но имеет важное отличие: она включает 53 языковых вложения, что позволяет модели эффективно работать с каждым поддерживаемым языком. Этот подход даёт возможность модели улавливать тонкости языков, даже если они имеют схожее написание или произношение. Модель включает в себя 500 миллионов параметров, что делает её одной из крупнейших многоязычных моделей в мире. XLSR-53 была обучена на огромном корпусе данных, включающем тексты и речь более чем на 53 языках. Её способность к межъязыковому пониманию особенно ценна для трансферного обучения. Это позволяет использовать модель, обученную на одном языке, для достижения высоких результатов на другом языке с минимальными затратами на дополнительное обучение.

Идентификация языка стало важной областью исследований в области обработки речи [10, 11, 12, 13], с различными моделями, введенными для повышения точности, особенно в многоязычных средах. В одном из таких исследований [14] были применены модели глубокого обучения для распознавания устной речи с использованием сверточных нейронных сетей (CNN) на спектрограммах необработанных аудиосигналов. Этот подход показал многообещающие результаты с точностью 98% на тестовом наборе данных, что демонстрирует эффективность CNN в извлечении специфичных для языка признаков из речевых данных. В этой работе [15] представлен многомасштабный подход к извлечению признаков для повышения точности идентификации языка (LID) в сложных акустических средах. Заменив базовую сеть извлечения признаков архитектурой SE-Res2Net-CBAM-BILSTM, исследование достигло значительного улучшения производительности. Эксперименты, проведенные на многоязычном наборе данных коктейльной вечеринки, продемонстрировали надежность модели, при этом точность достигла 97,6% для набора данных восточного языка и 75% для имитированного набора данных коктейльной вечеринки. Кроме того, исследование подчеркивает эффективность фокальной потери в улучшении производительности модели, особенно при обработке сценариев несбалансированных данных. Недавние достижения в

области обнаружения языка для казахского и русского языков также продемонстрировали эффективность подходов глубокого обучения. Другие работы [16, 17] использовали рекуррентные нейронные сети с долговременной краткосрочной памятью для различения казахского и русского языков в новостных данных. Их модель, обученная на сегментах длиной всего 2 секунды, достигла точности 86% для казахского языка, что подчеркивает потенциал моделей на основе LSTM в обработке языков с ограниченными ресурсами. Еще одно исследование [18] было сосредоточено на идентификации языка в комментариях казахстанских новостных платформ, где обычно используются казахский и русский языки, часто с переключением кода. Авторы предложили двухэтапную структуру, сочетающую неконтролируемую нормализацию и наивную байесовскую классификацию, а также модель глубокого обучения с использованием сетей LSTM для классификации текста. Их подход улучшил современные результаты по казахскому языку, подчеркнув эффективность сочетания традиционных методов и методов глубокого обучения для многоязычных и смешанных текстов. Эта работа особенно актуальна для продвижения языковой идентификации в сценариях с ограниченными ресурсами и переключением кода. Кроме того, использование i-векторов и x-векторов было широко распространенным подходом в системах идентификации языка и говорящего. Интеграция этих методов была исследована в этой работе [19]. Авторы продемонстрировали надежность систем i-векторов при применении к LID. Несмотря на их успех, ограничения в обработке коротких высказываний и шумной среды привели к дальнейшему развитию в этой области.

Для преодоления проблем языков с низкими ресурсами кросс-языковые модели, такие, как CLSR, стали более эффективным решением. XLSR основан на фреймворке Wav2Vec2, что позволяет совместно использовать речевые представления на нескольких языках, что делает его особенно мощным инструментом для LID в многоязычных условиях [9]. Целью данного исследования является демонстрация улучшений как точности, так и надежности путем тонкой настройки XLSR на различных наборах данных, включая недостаточно представленные языки, такие, как казахский.

3. МЕТОДОЛОГИЯ

В этом разделе рассматриваются наборы данных, используемые для идентификации языка, включая казахский язык. Также в нем рассматриваются методологии, используемые для разработки модели идентификации языка на основе тонкой настройки модели wav2vec2.

В основе модели лежит преобразователь-кодер, который обрабатывает квантованные признаки и фиксирует как локальные, так и глобальные закономерности в речи. Это позволяет модели обрабатывать последовательности различной длины и понимать контекст за пределами отдельных звуков. Кросс-языковая природа XLSR позволяет модели изучать общие представления нескольких языков, что делает ее идеальной архитектурой модели для задач LID. Окончательный вывод создается классификационной головкой, которая определяет наиболее вероятные последующие токены. Эта архитектура позволяет модели достигать высокой точности в идентификации языка, используя общие закономерности в языках с высоким и низким ресурсом.

Набор данных, используемый для этой модели идентификации языка, был составлен из нескольких источников, чтобы обеспечить всестороннее и разнообразное представление языков. Основной набор данных, VoxLingua107 [20], был выбран из-за его обширного охвата многоязычных данных в задачах идентификации языка. Для дальнейшего улучшения набора данных были получены дополнительные данные для русского и английского языков из набора данных Common Voice 17.0 [21]. Для казахского языка использовался Корпус казахской речи 2, разработанный исследовательской лабораторией ISSAI в Назарбаев Университете [22]. Набор данных охватывает шесть языков: казахский, русский, английский, итальянский, испанский и корейский. Эти языки были выбраны на основе их релевантности задаче. Были предприняты согласованные усилия, чтобы гарантировать, что итальянский, испанский и корейский языки

также были представлены в наборе данных для поддержания разнообразия и надежности в задаче идентификации языка.

Для обеспечения высококачественного ввода в модель идентификации языка были предприняты шаги по предобработке данных. Аудиофайлы из каждого набора данных были конвертированы в единый формат WAV с частотой дискретизации 16 кГц, чтобы обеспечить совместимость с моделью wav2vec2. Продолжительность аудиофайлов была приведена к диапазону от 5 до 30 секунд, чтобы избежать перекосов из-за чрезмерно коротких или длинных файлов. Структура набора данных была организована следующим образом: каждая папка соответствовала определенному языку, а внутри нее аудиофайлы и соответствующие текстовые аннотации были структурированы с использованием уникальных идентификаторов для упрощения обработки. Статистические данные по обучающим и тестовым наборам данных приведены в Таблицах 1 и 2.

Таблица 1 – Статистика по набору обучающих данных

Метка	Всего аудио	Общая продолжительность (s)	Средняя продолжительность (s)	Мин. продолжительность (s)	Макс. продолжительность (s)
kk	42 506	357 510.61	8.41	1.00	24.78
ru	24 969	295 473.74	11.83	1.52	20.00
es	25 439	243 771.40	9.58	1.15	20.00
it	3 181	30 211.77	9.50	2.00	20.00
en	2 677	29 124.43	10.88	1.97	19.98
ko	3 018	27 188.59	9.01	1.72	20.00
Общий	101 790	983 280.54	9.66	1.00	24.78

Таблица 2 – Статистика по тестовому набору данных

Метка	Всего аудио	Общая продолжительность (s)	Средняя продолжительность (s)	Мин. продолжительность (s)	Макс. продолжительность (s)
kk	10 568	89 112.31	8.43	1.00	24.58
ru	6 354	61 122.64	9.62	1.22	20.00
es	673	7 243.62	10.76	2.37	20.00
it	824	7 964.66	9.67	2.00	19.86
en	6 301	74 041.94	11.75	1.41	20.00
ko	731	6 849.41	9.37	1.81	19.76
Общий	25 448	246 334.57	9.68	1.00	24.58

Набор данных был разделен на обучающий и тестовый наборы, 80% выделено для обучения и 20% зарезервировано для тестирования. Разнообразие и размер набора данных сыграли решающую роль в обеспечении обучения модели на репрезентативном наборе языковых высказываний.

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Обучение проводилось с использованием оптимизатора AdamW со скоростью обучения от 3 до 5. Коэффициент разогрева, равный 0,1, был применен для постепенного увеличения скорости обучения по сравнению с начальными этапами, что важно для достижения стабильного обучения. Модель обучалась в течение 25 периодов времени на одном графическом процессоре, при этом размер пакета для каждого устройства составлял 64, а количество шагов накопления градиента - 4, что эффективно увеличивало размер пакета.

Для оптимизации использования памяти графического процессора и скорости обучения использовалось смешанное прецизионное обучение. Конвергенция обучения и потеря оценки во время обучения модели показаны на Рисунке 1.

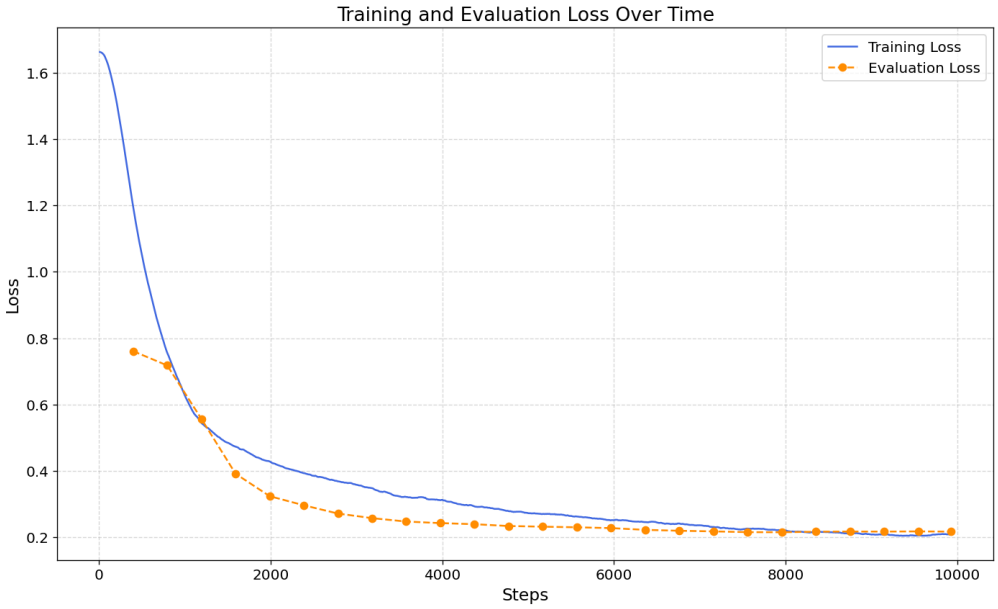

Рисунок 1 – Конвергенция обучения и потеря оценки во время обучения модели

В ходе процесса отслеживались как обучение, так и оценка. Потери обучения неуклонно снижались с течением времени, а потери оценки следовали аналогичной траектории, что указывает на эффективное обучение и хорошую сходимость модели. Ранняя остановка не потребовалась, поскольку не было никаких сигналов переобучения на протяжении 25 эпох. Оценка точно настроенной модели XLSR проводилась путем сравнения ее производительности с несколькими моделями Whisper (Tiny, Small и Large Whisper-3) [23] с использованием того же тестового набора данных. Основными метриками, использованными для сравнения, были точность, прецизионность, отзыв и оценка F1, которые были широко принятыми мерами для задач классификации, таких, как идентификация языка. Сравнение результатов, полученных с помощью моделей, представлено в Таблице 3.

Таблица 3 – Сравнение производительности моделей

Модель	**Точность (Accuracy)**	**Точность (Precision)**	**Полнота (Recall)**	**оценка F1**
Tiny Whisper	0.63	0.90	0.63	0.62
Small Whisper	0.76	0.92	0.76	0.79
Large Whisper	0.85	0.91	0.85	0.86
Fine-tuned Model	**0.93**	**0.93**	**0.93**	**0.93**

Tiny Whisper: несмотря на достижение высокой точности 0,9, она показала относительно низкую точность 0,63 и оценку F1 0,62. Это говорит о том, что, хотя модель могла хорошо предсказывать определенные языки, ее общая способность к полноте и обобщению была ограничена.

Small Whisper: эта модель продемонстрировала разумное улучшение по сравнению с Tiny Whisper с точностью 0,76 и оценкой F1 0,79. Значения полноты и точности были

сбалансированы, что указывает на лучшую производительность в определении более широкого диапазона языков. Однако она все еще отставала с точки зрения общей точности.

Large Whisper-3: как и ожидалось, модель Large Whisper-3 показала значительные улучшения как в точности (0,85), так и в оценке F1 (0,86). Эта модель обеспечила более сильные возможности идентификации языка, но уступила модели XLSR с точки зрения полноты и общей производительности по всем языкам.

Fine-tuned XLSR: Модель XLSR превзошла все другие модели по всем показателям, достигнув точности 0,93 и оценки F1 0,9293. Это указывает на то, что процесс тонкой настройки на разнообразном многоязычном наборе данных, особенно включая языки с низкими ресурсами, такие, как казахский, значительно улучшил способность модели обобщать между языками и обеспечивать высокую производительность в реальных задачах идентификации языка.

Для дальнейшего понимания возможностей модели XLSR был проведен анализ ошибок, чтобы выявить основные причины неверных классификаций. Анализ ошибок модели выявил основные причины неверных классификаций и возможные пути их устранения. Наибольшее количество ошибок возникало при идентификации языков с высокой степенью фонетической и морфологической схожести, таких как испанский и итальянский, шведский и датский, а также русский и казахский, особенно в случаях заимствованных слов. Значительная доля ошибок была связана с низким качеством данных, включая фоновые шумы, обрывки фраз и артефакты в аудиозаписях. Проблемы с классификацией также наблюдались при код-свичинге, когда модель некорректно интерпретировала смешанные фразы, например, на русском и казахском, что снижало точность. Для улучшения качества классификации рекомендуется усилить предварительную обработку данных, включая шумоподавление, добавить больше примеров код-свичинга в тренировочный набор, а также рассмотреть использование дополнительных языковых признаков и механизмы для анализа мультиязычного контекста.

Ключевое преимущество обученной модели XLSR заключается в её способности эффективно работать как с языками с высоким уровнем ресурсов, так и с низкоресурсными языками. В то время как модели, такие, как Whisper, демонстрировали хорошие результаты в условиях высокого уровня ресурсов, использование перекрестного языкового переноса в модели XLSR позволило ей добиться успеха даже с языками, для которых было доступно ограниченное количество данных для обучения, такими, как казахский. Это подчеркивает ценность использования общих представлений между языками, что является неотъемлемой особенностью архитектуры XLSR.

5. ВЫВОДЫ

В этой работе подчеркивается исключительная эффективность тонкой настройки модели XLSR для многоязычной идентификации языков с уделением особого внимания решению проблем, связанных с языками с ограниченными ресурсами. Благодаря стратегическому использованию межъязыкового трансферного обучения модель XLSR значительно превзошла традиционные модели, включая варианты Whisper, достигнув заметных улучшений как в точности, так и в показателях F1. Такая производительность демонстрирует замечательную способность модели справляться с различными лингвистическими сложностями, адаптироваться к изменчивости ресурсов и поддерживать надежность в широком спектре языковых задач. Полученные результаты подчеркивают преобразующую роль XLSR в преодолении разрыва между языками с высокой и низкой ресурсоемкостью, предлагая масштабируемое и эффективное решение для разработки передовых многоязычных систем обработки речи. Эта работа указывает на многообещающий путь к демократизации доступа к передовым языковым технологиям для недопредставленных языков по всему миру.

Вклад авторов

Умбет С. – Исследование, методология, курирование данных, формальный анализ, валидация, визуализация.

Кожирбаев Ж.М. – Концептуализация, курирование данных, формальный анализ, получение финансирования, исследование, методология, ресурсы. Написание - оригинальный черновик, написание - рецензирование и редактирование.

Список литературы

1 Niesler, T. R., Willett, D. Language identification and multilingual speech recognition using discriminatively trained acoustic models // Proceedings of Interspeech. - Pittsburgh, PA, USA, 2006. - P. 134-137.

2 Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in neural information processing systems. - 2020. - V. 33. - P. 12449-12460.

3 Song J., Ermon S. Multi-label contrastive predictive coding // Advances in Neural Information Processing Systems. - 2020. - V. 33. - P. 8161-8173.

4 Li S., Li L., Hong Q., Liu L. Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning // Proceedings of Interspeech. - Shanghai, China, 2020. - P. 5006-5010.

5 Schneider S., Baevski A., Collobert R., Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition // Proceedings of Interspeech. - Graz, Austria, 2019. - P. 3465-3469.

6 Baevski A., Schneider S., Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations // Proceedings of 8th International Conference on Learning Representations (ICLR). - Addis Ababa, Ethiopia, 2020. - P. 1-12.

7 Fan, Z., Li, M., Zhou, S., Xu, B. Exploring wav2vec 2.0 on Speaker Verification and Language Identification // Proceedings of Interspeech. - Brno, Czechia, 2021. - P. 1509-1513.

8 Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. - Minneapolis, Minnesota, USA, 2019. - P. 4171-4186.

9 Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. Unsupervised cross-lingual representation learning for speech recognition // Proceedings of Interspeech. - Brno, Czechia, 2021. - P. 2426-2430.

10 Kozhirbayev, Z., Islamgozhayev, T. Cascade speech translation for the Kazakh language // Applied Sciences. - 2023. -V. 13(15). -P. 8900.

11 Kozhirbayev, Z. Kazakh Speech Recognition: Wav2vec2. 0 vs. Whisper // Journal of Advances in Information Technology. - 2023. -V. 14(6). -P. 1382-1389.

12 Kozhirbayev, Z., Karabalayeva, M., Yessenbayev, Z. Spoken term detection for kazakh language // Proceedings of the 4-th International Conference on Computer Processing of Turkic Languages "TurkLang 2016". - Bishkek, 2016. -P. 47.

13 Kozhirbayev, Z., Yessenbayev, Z. Semantically expanded spoken term detection // IEEE Access. -2024. -V. 12. -P. 177844-177855.

14 Singh, G., Sharma, S., Kumar, V., Kaur, B., Bax, M., Masud, M. Spoken Language Identification Using Deep Learning // Computational Intelligence and Neuroscience. - 2021. - V.1. -P. 5123671.

15 Aysa, Z., Ablimit, M., Hamdulla, A. Multi-scale feature learning for language identification of overlapped speech // Applied Sciences. - 2023. - V.13(7). - P. 4235.

16 Kozhirbayev, Z., Yessenbayev, Z., Karabalayeva, M. Kazakh and Russian languages identification using long short-term memory recurrent neural networks // Proceedings of the 11th International Conference on Application of Information and Communication Technologies (AICT). – Moscow, Russia, 2017. -V. 1. -P. 1–5.

17 Kozhirbayev, Z., Yessenbayev, Z., Sharipbay A. Language identification in the spoken term detection system for the kazakh language in a multilinge environment // Journal of Mathematics, Mechanics and Computer Science. - 2019. -V. 96(4). -P. 88–98.

18 Kozhirbayev, Z., Yessenbayev, Z., Makazhanov, A. Document and word-level language identification for noisy user generated text // Proceedings of the 12th International Conference on Application of Information and Communication Technologies (AICT). - Almaty, Kazakhstan, 2018. - P. 1-4.

19 Shen, P., Lu, X., Li, S., Kawai, H. Conditional generative adversarial nets classifier for spoken language identification // Proceedings of Interspeech. - Stockholm, Sweden, 2017. – P. 2814-2818.

20 Valk, J., Alum?e, T. Voxlingua107: a dataset for spoken language recognition // Proceedings of IEEE Spoken Language Technology Workshop (SLT). - Shenzhen, China, 2021. - P. 652-658.

21 Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., Weber, G. Common Voice: A Massively-Multilingual Speech Corpus // Proceedings of the Twelfth Language Resources and Evaluation Conference. -Marseille, France, 2020. - P. 4218-4222.

22 Mussakhojayeva, S., Khassanov, Y., Varol, H. A. KSC2: An industrial-scale open-source Kazakh speech corpus // Proceedings of Interspeech. - Incheon, Korea, 2022. - P. 1367-1371.

23 Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. Robust speech recognition via large-scale weak supervision // Proceedings of the International conference on machine learning. – Hawaii, USA, 2023. - P. 28492-28518.

Қазақ тіліне арналған Wav2Vec2 моделін пайдалана отырып, ауызша сөйлеу тілін сәйкестендіру

С. Умбет [1] **, Ж.М. Кожирбаев** [2]
[1] Трир университеті, Университетсринг көшесі 15, 54296, Трир, Германия
[2] National Laboratory Astana, Қабанбай батыр даңғылы, 53, Астана, Қазақстан

Аңдатпа. Бұл зерттеу XLSR (Cross-Lingual Speech Recognition) WAV2VEC2 моделін қолдана отырып, ауызша тілді сәйкестендіру моделін әзірлеуді және дәл баптауды ұсынады. Алты тілді қамтитын бай және алуан түрлі деректер жиынтығында оқытылған, қазақ тілі сияқты ресурстары төмен тілдерге ерекше назар аудара отырып, модель көп тілді сөйлеуді танудың керемет мүмкіндіктерін көрсетеді. Кең бағалаудың арқасында дәл бапталған модель қолданыстағы эталондардан асып қана қоймайды, сонымен қатар басқа заманауи модельдерден, соның ішінде Whisper нұсқаларынан да асып түседі. F1 92,9% және 93% дәлдікпен жоғары нәтижеге қол жеткізген модель өзінің өнімділігін нақты көп тілді және аз ресурстық сценарийлерде көрсетеді. Бұл жұмыс сөйлеуді тану технологияларының дамуына айтарлықтай үлес қосады, әр түрлі тілдік ортада, әсіресе аз ұсынылған тілдік жағдайларда тілді анықтаудың сенімді шешімін ұсынады. Оның жетістігі wav2vec2 негізіндегі модельдердің ресурстары төмен көп тілді контекстерде сөйлеуді өңдеу жүйелерін жақсартудағы әлеуетін көрсетеді. Осы талдаудың қорытындылары қазақ тілі үшін оңтайландырылған сөйлеуді автоматты түрде танудың сенімді және тиімді жүйелерін әзірлеуге ықпал етуі мүмкін. Мұндай технологиялар әртүрлі салаларда, соның ішінде сөйлеуді мәтінге түрлендіруде, дауыстық көмекшілердің жұмысында және дауыстық байланыс құралдарында қолданылады.

Түйін сөздер: тілді сәйкестендіру, ауызша сөйлеу тілін сәйкестендіру, қазақ тілі, Wav2Vec2, XLSR.

Identification of the spoken language using the Wav2Vec2 model for the Kazakh language

S. Umbet [1] **, Zh.M. Kozhirbayev** [2]
[1] University of Trier, st. Universitetsring 15, 54296, Trier, Germany
[2] National Laboratory Astana, Kabanbay batyr ave. 53, Astana, Kazakhstan

Abstract. This study presents the development and fine-tuning of an oral language identification model using the XLSR (Cross-Lingual Speech Recognition) Wav2Vec2 variant. Trained on a rich and diverse dataset spanning six languages, with a particular focus on low-resource languages such as Kazakh, the model demonstrates remarkable capabilities in multilingual speech recognition. Thanks to extensive evaluation, the finely tuned model not only surpasses existing benchmarks, but also surpasses other modern models, including Whisper variants. Having achieved an impressive F1 score of 92.9% and an accuracy of 93%, the model demonstrates its performance in real multilingual and low-resource scenarios. This work makes a significant contribution to the development of speech recognition technologies by providing a reliable solution for language identification in various language environments, especially in underrepresented language settings. Its success highlights the potential of Wav2Vec2-based models in improving speech processing systems in low-resource multilingual contexts. The results of this analysis can contribute to the development of reliable and effective automatic speech recognition systems optimized for the Kazakh language. Such technologies will find applications in various fields, including speech-to-text conversion, voice assistants and voice communication tools

Keywords: language identification, spoken language identification, Kazakh language, Wav2Vec2, XLSR.

References

1 Niesler T. R., Willett D. Language identification and multilingual speech recognition using discriminatively trained acoustic models, Proceedings of Interspeech, Pittsburgh, PA, USA, 2006. P. 134-137.
2 Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems. 2020. Vol. 33. P. 12449-12460.
3 Song J., Ermon S. Multi-label contrastive predictive coding, Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 8161-8173.
4 Li S., Li L., Hong Q., Liu L. Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning, Proceedings of Interspeech, Shanghai, China, 2020. P. 5006-5010.
5 Schneider S., Baevski A., Collobert R., Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition, Proceedings of Interspeech, Graz, Austria, 2019. P. 3465-3469.
6 Baevski A., Schneider S., Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations, Proceedings of 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020. P. 1-12.
7 Fan Z., Li M., Zhou S., Xu B. Exploring wav2vec 2.0 on Speaker Verification and Language Identification, Proceedings of Interspeech, Brno, Czechia, 2021. P. 1509-1513.
8 Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, USA, 2019. P. 4171-4186.
9 Conneau A., Baevski A., Collobert R., Mohamed A., Auli M. Unsupervised cross-lingual representation learning for speech recognition, Proceedings of Interspeech, Brno, Czechia, 2021. P. 2426-2430.
10 Kozhirbayev Z., Islamgozhayev T. Cascade speech translation for the Kazakh language, Applied Sciences. 2023. Vol. 13(15). P. 8900.
11 Kozhirbayev Z. Kazakh Speech Recognition: Wav2vec2. 0 vs. Whisper, Journal of Advances in Information Technology. 2023. Vol. 14(6). P. 1382-1389.
12 Kozhirbayev Z., Karabalayeva M., Yessenbayev Z. Spoken term detection for kazakh language, Proceedings of the 4-th International Conference on Computer Processing of Turkic Languages “TurkLang 2016”, Bishkek, 2016. P. 47.
13 Kozhirbayev Z., Yessenbayev Z. Semantically expanded spoken term detection, IEEE Access. 2024. Vol. 12. P. 177844-177855.
14 Singh G., Sharma S., Kumar V., Kaur B., Bax M., Masud M. Spoken Language Identification Using Deep Learning, Computational Intelligence and Neuroscience. 2021. Vol.1. P. 5123671.
15 Aysa Z., Ablimit M., Hamdulla A. Multi-scale feature learning for language identification of overlapped speech, Applied Sciences. 2023. Vol.13(7). P. 4235.
16 Kozhirbayev Z., Yessenbayev Z., Karabalayeva M. Kazakh and Russian languages identification using long short-term memory recurrent neural networks, Proceedings of the 11th International Conference on Application of Information and Communication Technologies (AICT), Moscow, Russia, 2017. Vol. 1. P. 1–5.
17 Kozhirbayev Z., Yessenbayev Z., Sharipbay A. Language identification in the spoken term detection system for the kazakh language in a multilinge environment, Journal of Mathematics, Mechanics and Computer Science. 2019. Vol. 96(4). P. 88–98.
18 Kozhirbayev Z., Yessenbayev Z., Makazhanov A. Document and word-level language identification for noisy user generated text, Proceedings of the 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 2018. P. 1-4.
19 Shen P., Lu X., Li S., Kawai H. Conditional generative adversarial nets classifier for spoken language identification, Proceedings of Interspeech, Stockholm, Sweden, 2017. P. 2814-2818.
20 Valk J., Alum?e T. Voxlingua107: a dataset for spoken language recognition, Proceedings of IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 2021. P. 652-658.
21 Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F. M., Weber G. Common Voice: A Massively-Multilingual Speech Corpus, Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 2020. P. 4218-4222.
22 Mussakhojayeva S., Khassanov Y., Varol H. A. KSC2: An industrial-scale open-source Kazakh speech corpus, Proceedings of Interspeech, Incheon, Korea, 2022. P. 1367-1371.
23 Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision, Proceedings of the International conference on machine learning, Hawaii, USA, 2023. P. 28492-28518.

Сведения об авторах:

Санжар Умбет – магистрант по специальности "Наука о данных" в Триерском университете, ул. Университетсринг, 15, 54296, Трир, Германия.

Кожирбаев Жанибек Мамбеткаримович – PhD, старший научный сотрудник, National Laboratory Astana, пр. Кабанбай батыра, 53, Астана, Казахстан.

Sanzhar Umbet – masters student in Data Science at University of Trier, st. Universitetsring 15, 54296, Trier, Germany.

Zhanibek Kozhirbayev – – PhD, senior researcher, National Laboratory Astana, Kabanbay batyr ave. 53, Astana, Kazakhstan.

Поступила: 04.12.2024. После редакции: 23.01.2025.
Одобрена: 13.03.2025. Доступна онлайн: 31.03.2025.